On Features and Categories (CDT-20)

Luciano da Fontoura Costa
luciano@ifsc.usp.br

São Carlos Institute of Physics – DFCM/USP

February 14, 2020

Abstract

Along time, humans organized several entities of the real world into categories, as a means to summarize and abstract their properties while avoiding an explosion of labels that would be otherwise needed to identify every possible individual entity. The task of assigning categories to given entities, often represented in terms of a set of selected features, is called pattern recognition. This work discusses the interesting relationship between feature extraction/selection and respective assignment of categories (classification), especially regarding the type of mappings required for transforming entities into feature vectors, and then mapping these vectors into categories. The current work also focuses on the implications of the fact that supervised classification relies on pre-definition of the reference categories (e.g. prototypes) obtained by previous unsupervised pattern recognition, possibly involving different sets of features that are no longer accessible.

'What's in a name? That which we call a rose, by any other name would smell as sweet...'

William Shakespeare.

1 Introduction

Pinneapple... Watermelon... Lemon... Each word has an inherent ability to induce definite, palpable memories and feelings in the human mind. We read *pinneaple* and soon remember its yellow color, its remarkable shape, as well as its characteristic smell and taste. Ultimately, the power of words derive from their ability to almost instantly activate our memory, conveying definite meaning and a whole set of evocations.

It is therefore hardly surprising that so much has been studied and written about language. Yet, the close relationship between *linguistics* and *pattern recognition* (e.g. [1, 2]) is not so often realized or acknowledged. Interestingly, each word can be understood as a category – or even a model (e.g. [3, 4]) – in the sense of representing a group of real entities properties that are similar within the group, but which differ from entities belonging to other groups.

Equally important, is the fact that words were preceded by the conceptualization, by humans, of patterns

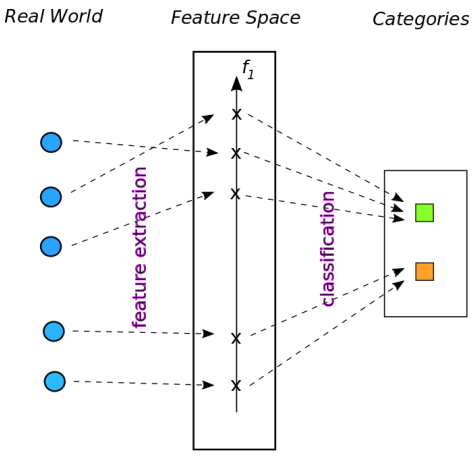

Figure 1: World entities (blue circles) are to be organized into categories. First, some of their properties – in this case the single feature f_1 – are extracted, implementing a mapping into a respective feature space. Observe that as much accurate and discriminative feature extraction is typically sought, hopefully yielding a bijective mapping from the entities into the feature space. Contrarily, a non-injective mapping then has to be applied in order to classify the points into categories (two in the case of this example).

of special relevance and importance. Such a representation of world entities in terms of respective categories allowed not only those entities to be effectively summarized, but also avoided the explosion of labels that would otherwise be necessary to identify all possible individual entities. Words appeared when it became necessary to communicate the human-defined categories by sound or graphically. However, the summarization and abstraction allowed by categories implies the impossibility to completely recover a specific individual entity from its category. Indeed, the use of adjectives and additional properties are often required for more accurate specification of entities.

The close relationship between linguistics and pattern recognition bears several important consequences. For instance, it suggests that the former can be modeled by using concepts and methods from the latter. At the same time, it allows the latter to be used as a means of implementing automated word and speech recognition, and even automated text generation. Moreover, given that much of our brain is oriented to language, pattern recognition-based approaches to language can ultimately contribute to better understanding of our brain.

On the other hand, it is also possible to employ concepts from linguistics and neuroscience in order to derive more effective pattern recognition methods. This is particularly important because most pattern recognition tasks ultimately refer to human concepts. For instance, if we want to recognize fruits, it is ultimately to the human conceptualization of these entities that we need to resource to and be compatible with. Even more specific applications not directly related to humans, such as automatically identifying car plate licenses, will at some point involve an interface with human cognition.

One particularly interesting issue which can potentially be better understood in the light of the above considerations concerns the relationship between entities *features* and *categories*.

First, it is important to realize that any object from the real world, and even abstractions, need to be quantified into a set of properties (also called features, measurements, attributes, etc.) prior to being handled by the nervous system or a computer. The action of taking measurements from a given entity, often called *feature extraction*, can be understood as a functional mapping from a domain (the real world) into a co-domain (a set of mathematical measurements). Typically, the adopted features are organized as a *feature vector* $\vec{f}$, so that each f_i, $i = 1, 2, \ldots, N$ corresponds to a particular measurement.

It is know from mathematics that a functional mapping is invertible if and only each entity in the domain is assigned to only one entity in the codomain (injection) and all entities in the codomain receive mappings from the domain (surjection). In this case, the mapping is said to be *bijective*. It is interesting to consider the assignment of categories according to these properties. Ideally, it would be interesting to have a bijective relationship between real-world entities and their respective categories, which would ensure each entity to be more accurately represented, contributing to the chances of being uniquely identified and eventually recovered.

However, because entities always have, often minute, properties that distinguish them, it is virtually impossible to achieve a bijective mapping between entities and their respective feature representation, which can imply in several entities being mapped into the same set of features values. Therefore, features should be selected with basis on their discriminative power between the entities of interest, in order to reduce the mapping ambiguity. For instance, the length of the beak of a bird provides a potentially discriminating feature between those animals. Typically, the chosen set of features should be as small as possible while allowing the pattern recognition problem in question to be solved in a satisfactory manner.

There are two main ways in which a more accurate mapping can be obtained: (i) by specifying more properties; and (ii) by restricting the domain. For instance, we can refer to a medium-sized tree with small leaves and long roots. Yet, even so, we will not achieve a bijective mapping if all trees in the world need to be considered. The fact that humans usually get along with single words is because the domain is restricted to the more immediate context. For instance, in a given classroom, when one says *teacher*, the subject is completely specified as corresponding to the person who is currently teaching in that room.

Then, we have the definition of the categories themselves, in this work called *classification* or *categorization*, which is based on the general principle that entities in a same category will have similar properties while differing from entities in other categories. Two main types of pattern recognition are often identified: (i) *supervised*, (e.g. [5] in which prototypes or specifications of each of the categories are available; and (ii) *unsupervised* (also called *clustering*, e.g. [6, 7, 8]), when nothing *a priori* is known about the categories, not even their number.

Consider the unsupervised situation illustrated in Figure 1. Here, we have 5 entities from the real world being mapped into respective feature space representations. For simplicity's sake, an one-dimensional feature space is considered, but the following discussion applies immediately to higher dimensions. In this particular example, the feature extraction stage implied the upper three entities to be mapped into similar, but not identical, values along the f_1 axis, while the other two entities were mapped into

another group of similar values along the feature space. When a clustering method is applied on the obtained feature representation, two categories are likely to emerge, as illustrated in the figure.

As expected, the objects within each of the obtained categories will share similar feature values, while being different from the entities in the other category. It is interesting to observe that the two subsequent stages of feature extraction and classification implies different demands on the features. Indeed, the representation of the original entities in terms of feature values should, in principle, be performed so as to maximize the preservation of properties capable of discriminating the entities, implying in a more accurate and complete mapping. However, at the following classification stage, the smaller differences between objects in a same group need to be disregarded, so that they can be merged, through a non-injective mapping, into respective categories! Therefore, in a sense, the stages of feature representation and classification can be understood as exerting contradicting demands on the considered set of features. Yet, the net result is to preserve discriminability between the entities of a same abstracted class (e.g. [9]).

One interesting problem involving the relationship between features and categories regards trying to decide, given a pattern recognition problem, if the features come before defining the categories, or if pre-defined categories should provide the basis for selecting the best features. Essentially, we have a kind of chick-and-egg problem.

Let's consider the two following examples: (1) given a set of handwritten texts from a recently discovered ancient civilization, unlike any known counterparts, one is required to infer its basic categories of characters; (2) one needs to implement a pattern recognition system to recognize apples and bananas. As we will see each of these problems underlies a specific relationship between feature extraction/selection and categorization.

Problem (1) above corresponds to an example of *unsupervised* pattern recognition, or clustering, given that nothing is previously known about the characters. In this case, no categories prototypes or previous characterization of the possible characters are available, one has no other choice than to extract/select a set of effective measurements of each separated symbol and try to find respective groups. It soon becomes clear that:

> **Unsupervised** pattern recognition (clustering) can only proceed *from features.*

The selection of the possible features is particularly critical, as it directly impacts the obtained categories. For instance, little discriminative features will tend to blur the categories. Typically, feature selection can be helped by a previous visual inspection of the entities, performed in order to identify some possible discriminative properties. It is also possible to perform some type of automated feature selection, for instance by systematically trying several possible combinations. Ultimately, however, *the obtained categories will be a consequence of the adopted features and clustering method.*

Problem (2) provides one example of *supervised* pattern recognition: we have access not only to many prototypes but also have a good idea about the typical characteristics of apples and bananas. At least two approaches are therefore possible: (a) the specification of the properties (features) of each of these types of fruits is elaborated and used for classification (e.g. bananas have lengths never achieved by apples); and (b) the fruits to be classified are compared to pre-specified prototypes representative of each of the two considered categories. It follows that:

> **Supervised** pattern recognition can proceed *from features or categories – but wait...*

The case in which the classes are specified in terms of specific values of given features corresponds to the situation of supervised classification proceeding from a set of selected features. The second case, in which prototypes are available, provides an example of supervised pattern recognition following from categories.

The case of supervised classification proceeding from previously defined categories deserves further attention, as it is more complex than it may seem at first. For one reason, the quantitative comparison between the entities to be classified and the prototypes *has to be performed in terms of a metric space underlain by a set of respective features.* As a consequence, it is necessary to map both the entities to be classified, as well as the prototypes, into respective feature spaces before any objective comparison can be performed between them.

The important point here is that every previous determination of the categories had to be first performed as a clustering problem, based on a respective selection of features that may no longer be available. This is the case of the categories of apples and bananas, which was performed by humans, along a long time, based on a selection of features that was subjective and which are not accessible (they depended on specific neuronal activity considering several types of stimuli). As matter of fact, the set of features selected by each human being while identifying apples and bananas is by no means identical, and can change along time and space.

Figure 2 illustrates the situation of performing supervised classification with basis on predefined categories (prototypes) obtained by a previous clustering approach.

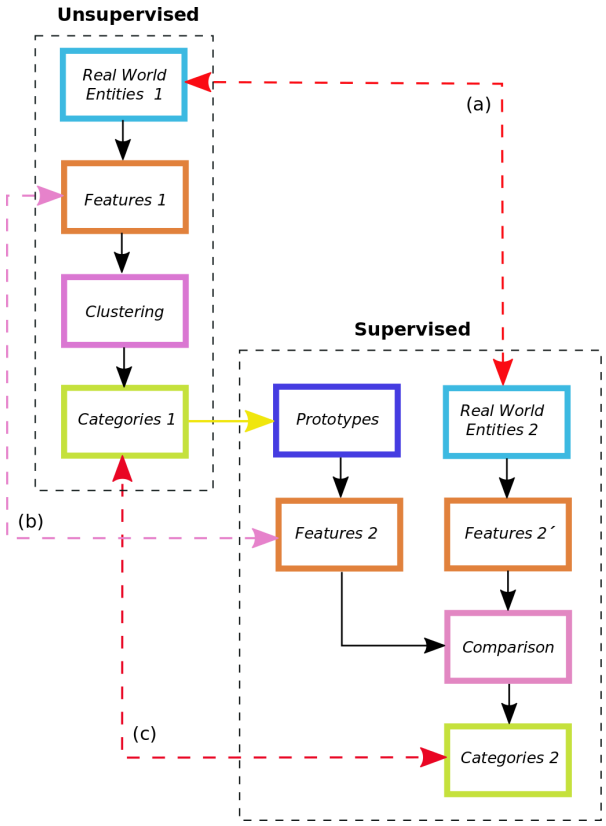

Figure 2: *Supervised classification presupposes clustering.* Entities from the real world have a specific set of features (1) extracted, giving rise, through clustering, to respective categories. For instance, we have the definition of apples and bananas by some human individual. Some objects classified into the obtained categories are then selected as prototypes in a supervised pattern recognition problem. A set of features (2) is selected that is unlikely to be identical to the set adopted in the previous clustering, and used for comparison between new real-world entities and the prototypes. The obtained categories are poised not to be fully congruent (c) with the previous clustering procedure because the original entities (a) and selected features (b) are unlikely to be the same. The dashed arrows indicate likely discrepancies between the two classification approaches.

In a preliminary unsupervised classification stage, several entities from the real world were mapped into categories with basis on the values of a set of selected features (1). For instance, this stage could correspond to the consolidation of the categories corresponding to apples and bananas by humans beings along time. In this way, though the categories are generally accepted and shared to a good extent, the features adopted for this classification are not accessible and likely to vary from individual to individual, and along time and space. Then, someone is required to implement a supervised pattern recognition system taking some fruits of each category as prototypes. These entities are mapped into a set of features (2) that are unlikely to be identical to those used before, allowing the comparison between the entities to be classified and the available prototypes. The distinct set of features implies that the obtained classification is very unlikely to be fully congruent with the previous clustering approach used to define the established categories and selected prototypes. We can therefore suggest that:

> Even when proceeding from predefined categories (e.g. prototypes), supervised pattern recognition still requires and relies on feature extraction/selection.

As a matter of fact, the framework identified in Figure 2 can yield further reaching implications. First, since the two selections of features are almost certainly different, and given that the entities are represented in terms of the respective selected features, even if the entities fed into the unsupervised and supervised stages are identical, they will be, to a good extent, *treated as different entities* along the respective classifications! In other words, the unsupervised and supervised pattern recognition systems in Figure 2 can be understood as not dealing with the same entity types. As a possible consequence, the obtained classification will refer more to the prototypes are represented by the feature set (2) than by to the original categories obtained by considering the feature set (1). These interesting results can be summarized as

> It is often difficult to implement supervised classification that is fully congruent with the original categories, unless identical feature sets and classification methods are adopted.

All in all, in this work we stressed the strong interrelationship between features and categories, with emphasis on the different types of mappings required for obtaining these respective objects, as well as on the fact that

supervised classification entails preliminary unsupervised pattern recognition possibly adopting distinct sets of features. It is hoped that the reported discussion can contribute not only to a more substantive appreciation of the intricacies and challenges of theoretical and applied aspects of pattern recognition, but also pave the way to devising enhanced respective concepts and methods.

Acknowledgments.

Luciano da F. Costa thanks CNPq (grant no. 307085/2018-0) for sponsorship. This work has benefited from FAPESP grant 15/22308-2.

References

[1] K. Koutrombas and S. Theodoridis. *Pattern Recognition*. Academic Press, 2008.

[2] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interscience, 2000.

[3] L. da F. Costa. Modeling: The human approach to science. Researchgate, 2019. `https://www.researchgate.net/publication/333389500_Modeling_The_Human_Approach_to_Science_CDT-8`. Online; accessed 03-June-2019.

[4] L. da F. Costa. Statistical modeling. `https://www.researchgate.net/publication/334726352_Statistical_Modeling_CDT-13`, 2019. [Online; accessed 22-Dec-2019.].

[5] D. R. Amancio, C. H. Comin, D. Casanova, G. Travieso, O. M. Bruno, F. A. Rodrigues, and L. da F. Costa. A systematic comparison of supervised classifiers. *PLOS One*, 2014.

[6] M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. da F. Costa, and F. A. Rodrigues. Clustering algorithms: A comparative approach. *PLOS One*, 14, 2019.

[7] U. v. Luxburg, R. C. Williamson, and I. Guyon. Clustering: Science or art? In *JMLR: Workshop and Conference Proceedings*, pages 65–79, 2012.

[8] C. Hennig. What are the true clusters? *Pattern Recognition Letters*, 64:53–62, 2015.

[9] C. H. Comin, F. Nascimento, and L. da F. Costa. A framework for evaluating complex networks measurements. *Europhysics Letters*, 110:68002, 2015.

Costa's Didactic Texts – CDTs

CDTs intend to be a halfway point between a formal scientific article and a dissemination text in the sense that they: (i) explain and illustrate concepts in a more informal, graphical and accessible way than the typical scientific article; and (ii) provide more in-depth mathematical developments than a more traditional dissemination work.

It is hoped that CDTs can also incorporate new insights and analogies concerning the reported concepts and methods. We hope these characteristics will contribute to making CDTs interesting both to beginners as well as to more senior researchers.

Each CDT focuses on a limited set of interrelated concepts. Though attempting to be relatively self-contained, CDTs also aim at being relatively short. Links to related material are provided in order to complement the covered subjects.

Observe that CDTs, which come with absolutely no warranty, are non distributable and for non-commercial use only.

The complete set of CDTs can be found at: `https://www.researchgate.net/project/Costas-Didactic-Texts-CDTs`.